METHOD OF FACE INDEXING FOR EFFICIENT BROWSING AND SEARCHING OF PEOPLE IN VIDEO

FIELD OF THE INVENTION

The present invention relates to a method and apparatus for generating indices of classes of objects appearing in a collection of images, particularly in a sequence of video frames, for use in searching or browsing for particular members of the respective class. The invention is particularly useful for indexing human faces in order to produce an index for face recognition tasks in querying operations.

DESCRIPTION OF THE RELATED ART

The amount of video data stored in multimedia libraries grows very rapidly which makes searching a time consuming task. Both time and storage requirements can be reduced by creating a compact representation of the video footage in the form of key-frames, that is a subset of the original video frames which are used as a representation for these original video frames. The prior art focuses on key-frame extraction as basic primitives in the representation of video for browsing and searching applications.

A system for video browsing and searching, based on key-frames, is depicted in Fig. 1A. A video image sequence is inputted from a video feed module 110. The video feed may be a live program or recorded on tape. Analog video is digitized in video digitizer 112. Optionally, the system may receive digital representation, such as Motion JPEG or MPEG directly. A user interface console 111 is used to select program and digitization parameters as well as to control key-frame selection. A key-frame selection module 113 receives the digitized video image sequence. Key-frames can be selected at scene transitions by detecting cuts and gradual transitions such as dissolves. This coarse segmentation into shots can be refined by selecting additional key-frames in a process of tracking changes in the visual appearance of the video along the shot. A feature extraction module 114 processes key-frames as

well as non key-frame video data to compute key-frames characteristic data. This data is stored in the key-frames characteristic data store 115 and is accessed by a video search engine 116 in order to answer queries generated by the browser-searcher interface 117. Such queries may relate to content attributes of the video data, such as color, texture, motion and others. Key-frame data can also be accessed directly by the browser 117. In browsing, a user may review key-frames instead of the original video, thus reducing storage and bandwidth requirements.

In an edited video program, the editor switches between different scenes. Thus, a certain collection of M video shots may consist only of N<M different scenes such that at least one scene spans more than one shot. The prior art describes how to cluster such shots based on their similarity of appearance, utilizing low-level features. It has become standard practice to extract features such as color, texture and motion cues, together with a set of distance metrics, and then utilize the distance metrics in the related feature spaces (respectively) for determining the similarity between key-frames of the video contents or shots. In this scenario, the video content is limited to the definition in the low-level feature space. What is missing in the prior art is the automatic extraction of high-level object-related information from the video during the indexing phase, so as to facilitate future searching applications.

In current systems for browsing and automatic searching, which are based on key-frames, the key-frames extraction and the automatic searching are separate processes. Combining the processes in a unified framework means taking into account high-level user-queries (in search mode), during the indexing phase. Spending more effort in a more intelligent indexing process proves beneficial in a short turn around rate in the searching process.

In automatic searching of video data by content, detecting and recognizing faces are of primary importance for many application domains such as news. The prior art describes methods for face detection and recognition in still images and in video image sequences.

A prior art method of face detection and recognition in video is depicted in Fig 1B. A face detection module (122) operates on a set of frames from the

input video image sequence. This set of frames may consist of the entire image sequence if the probability of detection is of primary importance. However, the face content in the video does not change with every frame. Therefore, to save computational resources a subset of the original sequence may be utilized. Such a subset may be derived by decimating the sequence in time by a fixed factor or by selecting key-frames by a key-frame extraction module 121. The detected faces are stored in the face detection data store 123.

For each detected face, face features are extracted (124), where the features can be facial feature templates, geometrical constraints, and global facial characteristics, such as eigen-features and features of other known algorithms in the art. The face representation can be compared to a currently awaiting search query, or to a predefined face database; alternatively, it can be stored in a face feature database (125) for future use. By comparing face characteristic data from the database or from a user-define query, with the face characteristic data extracted from the video, the identity of people in the video can be established. This is done by the face recognition module 126 and recorded in the video face recognition report 127 with the associated confidence factor.

Several algorithms for face recognition are described in the prior art. One prior art method uses a set of geometrical features, such as nose width and length, mouth position and chin shape; Another prior art method is based on template matching. One particular method represents the query and the detected faces as a combination of eigen-faces.

Co-pending Application No. PCT/IL99/00169 by the same assignee, entitled "Method of Selecting Key-Frames from a Video Sequence", describes a method of key-frame extraction by post-processing the key-frame set so as to optimize the set for face recognition. A face detection algorithm is applied to the key-frames; and in the case of a possible face being detected, the position of that key-frame along the time axis may be modified to allow a locally better view of the face. That application does not teach how to link between different views of the same person or how to globally optimize the different retained views of that person.

FIG. 1C shows a simple sequence of video scenes and the associated face content, or lack of it. In this example, some people appear in several scenes. Additionally, some scenes have more than one person depicted. FIG. 1D depicts the results of a sequential face-indexing scheme such as the one depicted in FIG. 1B. Clearly this representation provides only a partial, highly redundant description of the face content of the video scenes.

In a dynamic scene, a person may be visible for only a part, or for several parts, of the scene. In a scene and across the scenes, a person generally has many redundant views, but also several different views. In such a situation it is desirable to prune redundant views, and also to increase the robustness by comparing the user-defined query against all different views of the same person.

Also during a video segment, a person may go from a position where the person can be detected to a position where the person is visible but cannot be detected by automatic processing. In several applications it is useful to report the full segment of visibility for each recognized person.

SUMMARY OF THE INVENTION

According to one broad aspect of the present invention, there is provided a method of generating an index of at least one class of objects appearing in a collection of images for use in searching or browsing for particular members of the class, comprising: processing the collection of images to extract therefrom features characteristic of the class of objects; and grouping the images in groups according to extracted features helpful in identifying individual members of the class of objects.

According to a preferred embodiment of the invention described below, the collection of the images represents a sequence of video frames, the class of objects is human faces, and the grouping forms face tracks of contiguous frames, each track being identified by the starting and ending frames and containing face regions.

According to another aspect of the present invention, there is a provided a method of generating an index of human faces appearing in a sequence of video frames, comprising: processing the sequence of video frames to extract therefrom facial features; and grouping the video frames to produce face tracks of contiguous frames, each face track being identified by the starting and ending frames in the track and each face track containing face data characteristic of an individual face.

According to a still further aspect of the present invention, there is provided a method of generating an index of at least one class of objects appearing in a collection of images to aid in browsing or searching for individual members of the class, comprising: processing the collection of images to generate an index of features characteristic of the class of objects; and annotating the index with annotations.

The invention also provides a method of processing a sequence of video frames having a video track and an audio track, to generate a speech annotated face index associated with speakers, comprising: generating a face index from the video track; generating a transcription of the audio track; and aligning the transcription with the face index.

The invention further provides a method of processing a video track having face segments (including parts of face segments) to label such segments as talking or non-talking, comprising: tracking the face segments in the video track; detecting segments having mouth motion; and estimating from the detected segments, those having talking mouth motion vs. non-talking mouth motion.

According to a further aspect of the present invention, there is also provided a method of annotating a sequence of video frames, comprising: processing the sequence of video frames to generate a face index; and attaching a description to at least one entry in the face index.

The invention further provides a method of processing a sequence of video frames having a video track and an audio track, comprising: extracting from the video track, face segments representing human faces, and producing a face track for each individual face; extracting audio segments from the audio track; fitting a model based on a set of the audio segments corresponding to the individual face of a face track; and associating the model with the face track of the corresponding individual face.

According to another aspect of the present invention, there is provided a method of searching a collection of images for individual members of a class of objects, comprising: processing the collection of images to generate an index of features characteristic of the class of objects; and searching the index for individual members of the class of objects.

According to yet a further aspect of the present invention, there is a provided apparatus for generating an index of a class of objects appearing in a collection of images for use in searching or browsing for particular members of the class, comprising: a processor for processing the collection of images to extract therefrom features characteristic of the objects, and for outputting therefrom indexing data with respect to the features; a user interface for selecting the features to be extracted to enable searching for and identifying individual members of the class of objects; and a store for storing the index data outputted from the processor in groups linked according to the features selected for extraction.

As described more particularly below, the invention is particularly applicable for parsing a video stream into face/no face segments, indexing and logging the facial content, and utilizing the indexed content as an intelligent facial database for future facial content queries of the video data. The invention may use a high-level visual module in the indexing of a video stream, specifically, based on human facial information. Preferably, the invention also uses audio information as part of the indexing process and merges the audio and the video.

While the invention is particularly useful for indexing facial content in order to enable face recognition, the invention could also be used for indexing the characteristics of other classes of objects, such as billboards, logos, overlayed text (text added on to a video frame), geographical sites, etc. to enable recognition of individual members of such classes of objects, at an increased speed and/or probability of recognition.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A describes a prior art method of searching in video by key-frame selection and feature extraction.

FIG. 1B describes a prior art method of face recognition.

FIG. 1C describes a sample sequence of video content.

FIG. 1D presents the results of face detection applied to the sequence of FIG. 1C, organized as a sequential index.

FIG. 2 describes the browsing and searching system with face indexing, as introduced in this invention.

FIG. 3 depicts the face index.

FIG. 4 depicts a preferred embodiment of a face track structure.

FIG. 5 presents a sample face index generated by a particular embodiment of the present invention for the example of FIG. 1C.

FIG. 6 shows a set of face characteristic views selected from a face track, as taught by the present invention.

FIG. 7 provides an overview of the process of generating a face track from a video image sequence and extracting the associated characteristic data from it.

FIG. 8 shows a particular embodiment of a face tracking module.

FIG. 9 shows a particular embodiment for selecting the set of face characteristic views subject to self-similarity criteria.

FIG. 10a illustrates a template as the bounding box around a face region.

FIG. 10b illustrates the extraction of geometrical features from a face region.

FIG. 11 displays the vector store of the face characteristic data.

FIG. 12 shows the extraction of audio data corresponding timewise to the video data.

FIGS. 13 and 14 show how to combine the audio track with the face index to create an audio-visual index.

FIG. 15 describes the linking and information merging stage as part of the face indexing process.

FIG. 16 presents a sample video annotation generated by a particular embodiment of the present invention for the example of FIG. 1C.

FIG. 17 is a flow chart describing the method of searching an index and retrieving related video frames.

DETAILED DESCRIPTION OF THE INVENTION

As indicated above, a main purpose of the present invention is to provide a method of generating an index of a class of objects, such as a face object, appearing in a collection of images, such as a sequence of video frames, for use in searching or browsing for a particular member of the class, such as a specific individual's face (e.g., "Clinton"). The described method can be utilized for indexing various classes of objects (both visual objects and audio signatures) as long as the class of objects has a set of visual and/or audio characteristics that may be a-priori defined and modeled, so as to be used for automatic detection of objects within an input video stream, using video and audio analysis tools, as described herein.

According to the present invention an input video stream is automatically parsed into segments that contain the class of objects of interest, segments representative of particular members of said class of objects are grouped, and an index or index store is generated. The indexed content may be used for any future browsing and for efficient searching of any particular object. Some examples of classes of objects and particular members of the class include: face objects, and "Clinton"; logo objects and the "EuroSport" logo, text objects (automatic detection of words in the video stream) and "economy".

The following description relates primarily to the case of face indexing; however, the methods taught herein are general and can be readily applied to include indexing for browsing and searching of any class of objects as described above.

A system for video browsing and searching of face content in accordance with the present invention is depicted in FIG. 2. A video image sequence is inputted from a video feed module (210). The video feed may be a live program or recorded on tape. Analog video is digitized in a video digitizer (215). Optionally, the system may receive the digital representation directly. The video source, the program selection and digitization parameters, and the face-indexing selection parameters are all controlled by the user from an interface console (230). A subset of the video sequence is inputted into an

indexing module (220). The computed face index data is stored in the face index data store (250). Preferably the face-index data is edited by a face index editor to correct several types of errors that may occur during automatic indexing. Such errors can originate from false face detection, that is identifying non-face regions as faces. An additional form of error is over-segmentation of a particular face: two or more instances of the same face fail to be linked between appearances and thus generate two or more index entries. These errors are handled reasonably by any recognition scheme: false faces will generally not be recognized and over-segmentation will result in somewhat additional processing time and possibly reduced robustness. However, applications in which the generated index is queried frequently, it is generally cost-effective to have an editor review the graphical representation of the face index, to delete false alarms, and to merge entries originating from the same person.

Preferably, the editor annotates the face index store by specifying the name of the person, or by linking the face appearance and another database entity. This embodiment provides a method of semi-automatic annotation of video by first generating a face index and then manually annotating only the index entries. The annotation becomes immediately linked to all tracks in the video, which correspond to the specific index entry. Since the number of frames where a specific person is visible is much larger than the number of index entries, and since the spatial information, (that is the location within the frame) is readily available, a major effort saving is achieved. A more elaborate description of utilizing the face index store for annotation is provided below.

The face index can be utilized for efficient browsing and searching. Browsing can be effected via a browser-searcher module (240), and searching can be effected via a video search engine (260), both accessing the index in order to process face-based queries. An extended description of the browsing and searching operational modalities is provided below.

FIG. 3 illustrates one possible structure of the face index 250, which is provided with fields arranged in vertical columns and horizontal rows. As will be described below, each face entry (Fn), is associated with a plurality of tracks (Tn), along with representative data per track. A face track or segment is a

contiguous sequence of video frames in which a particular face appears. Each face track is depicted as a separate horizontal row in the index to include the information associated with the respective track as inputted into the face index store 250.

Generating an index as in FIG. 3 involves several processes. First, the input video is processed to automatically detect and extract objects of the predefined object category. Processing involves utilizing the set of visual and audio characteristics that are a-priori defined and modeled using video and audio analysis tools as described herein. A second major process involves grouping of the information in groups according to extracted features helpful in identifying individual members of a class. Grouping involves the generation of a track of contiguous frames (Tn), for each individual member of the class of objects to be identified. Grouping may also involve the merging of frames that contain similar members of a class of objects as indicated in associating different tracks (Tn) to a particular face entry (Fn).

A face track or segment is represented by a start frame, Fs, and an end frame, Fe. Face coordinates are associated with each frame in the track {X(Fs),...,X(Fe)}. In one embodiment, such coordinates could be the bounding-box coordinates of the face, as shown in FIG. 10a. In addition, the face track data structure includes:

- Face Characteristic Views, which are the visually distinct appearances of the face in the track.
- Face Frontal Views, which are the identified frontal views within the set of Face Characteristic Views; frontal views have better chances of being recognized properly.
- Face Characteristic Data, which are attributes computed from the face image data and stored for later comparison with the corresponding attributes extracted from the query image; such data can include audio characteristic data that can be associated with the face track.

FIG. 4 shows a preferred embodiment of a face track data structure.

FIG. 5 depicts a sample face index that is generated in module 250, FIG. 2, according to a particular embodiment of the invention for the example

depicted in FIG. 1C. The first detected face appearance is inputted as the first entry in the face index store (labeled as face A). The starting frame is inputted as fs=43. The end frame for the face track is frame fe=76. The track is inputted as a separate row (A1) in the data store, including the start and end frame indices, as well as the spatial coordinates of the face location in each frame of the sequence. In frames 77 thru 143, two faces are detected. One of them is found similar to face A, such that the face track is linked to face A entry in the face index store, indicated as a separate row, A2. The second face is a new detected face, initiating a new entry in the face index (labeled as face B), and the track information is inputted as a separate row, B1. The process is repeated for new face C, etc. It is to be noted that without an annotation process, face entries in the data store are labeled randomly (A,B,...), the main focus is to differentiate among them as separate entities. Annotation can be added to an index entry (right-most field). It can include a label (name of person), a description (affiliation, visual features, etc); it may be input manually or input automatically, as will be described in detail below.

FIG. 6 shows a set of face characteristic views selected from a face track in accordance with the present invention; a star denotes a face frontal view which may be selected as taught by the present invention.

FIG. 7 shows an overview of the process of generating a face track from a video image sequence, and extracting the associated characteristic data. The processing steps can be initiated at each frame of the video or more likely at a sub-set of the video frames, selected by module (710). The sub-set can be obtained by even sampling or by a key-frame selection process. Two process modules applied to the frames: a face detection module (720) and a face tracking module (730). In the face detection module (720), each frame is analyzed to find face-like regions. Face detection algorithms as known in the prior art may be utilized. Preferably, the detection method as taught by prior art, locates facial features in addition to the bounding box of the entire face region. Such features generally consist of eye features and additionally mouth features and nose features.

In the face tracking module (730), each set of consecutive frames (I,I+1) is analyzed to find correlation between faces in the two frames. Correlation based tracking is described below with respect to FIG. 8.

The face detection results and the face tracking results are merged and analyzed in module 740. A decision is made in logic block 750 whether to make a new entry insertion to the face track structure, and/or whether to augment an existing entry. The decision logic of block 750 may include the following set of rules. A current detected face in frame I+1 is defined as a new face entry if either no face was detected in the same spatial locality in frame I, or a face that did appear in the local neighborhood in frame I, is substantially non-similar in appearance and attributes. Once a detected face is declared as a new face entry, a face track entry is added to the face index (FIG. 3) and a new face track structure is initiated with the starting frame index. Initially, all detected faces are inputted into the face index, with each also initializing a face track structure.

In a second case, the current detected face region is part of an existing track. In one embodiment, correlation-based checking is pursued for a similarity check between the new face region and the faces in the previous frame. If a track is found, the existing track entry is augmented with information from the current face region, and the track index is updated to include current frame.

A third case should also be considered in which a track exists from the previous frame, I, yet the detection module (720) does not detect the face in the current frame, I+1. Rather, by using the track information, the tracking module (730) is able to find a face-like region that correlates with the existing track. Such a case could occur for example when a person is slowly rotating his head. The rotation angle may be large enough such that the face is no longer detectable as a face. Using the track information, however, the face region in the previous frame may guide the search for the face region in the current face, and a correlation may be found between the smoothly transitioned viewpoints of the face. In this case, the existing track entry is augmented with information from the current face region, and the face track index is updated to include the current frame.

Once a face track structure is terminated as described below with respect to FIG. 8, the characterizing sets are selected, and characterizing data is extracted, as indicated by block 760 and as described more particularly below. The terminated face track structure is taken out of the face track store and inserted and merged into the face index store (block 770).

FIG. 8 describes a particular embodiment of the face tracking module 730 effected via eye tracking. Initialized by a detection event at frame K, a defined tracking reference frame, R, is set to K (810), and the current frame, J, is set to the following frame in the sequence (820). Based on the location of features in the reference frame, R, a prediction or estimate can be made as to the most probable location of the same features in the current frame, J (830). For example, the location of the features can be estimated as the same location in the previous frame. Alternatively, trajectory estimates may be used, specifically when motion of the features is detected.

Utilizing the location estimates for features in frame R and corresponding features in frame J, a similarity transformation can be derived (840) as the mathematical definition of the disparity between the corresponding features. This transformation may be used to update R as a modified frame R1, such that any differences (e.g. zoom, rotation) between the new reference frame R1 and the current frame, J, are reduced (850).

Correlation matching may be used for the final feature search based on a best match search between the features in the two frames (860). For example, such a matching may involve the correlation of a window surrounding the features (e.g. eyes) in frame R1 with several windows at the locality of the predicted feature locations in frame J, and selecting the highest correlation score.

In case of a correlation score that is high, the detected face track is continued. A verification step may be included (870) enabling an update of the reference frame. The reference frame may be updated to eliminate the possibility of a continuous reduction in the correlation score between the reference frame and the new input frame (for example for a slowly rotating face), that may lead to a premature termination of the track. Thus, if consecutive

frames have a high correlation score yet the correlation score between the input frame and the reference frame is below a threshold, the reference frame is updated. In such a scenario, the reference frame R is updated to be in fact the current frame J. The iteration is completed, and the next frame in the sequence is analyzed in the next loop of the tracking process (back to 820). In case of no update, the reference frame remains unchanged in the next loop.

In case of a correlation score that is below an acceptable threshold, no match is found, and the end of a track is declared. The state of a track end is acknowledged as a track termination step. It leads to the extraction of feature sets representative of the track.

FIG. 9 shows a preferred embodiment for selecting the Face-Characteristic-Views subject to self-similarity criteria: The process starts with the start frame of a track entering the set, *C*, of Face-Characteristic-Views (905). Start frame, I, is taken as a reference frame. Next frame is loaded as frame K (910). Given the currently selected reference frame, I, the consecutive frame, K, is compared against I. In a procedure similar to the one described with respect to FIG. 8, the correspondence between facial features is used to solve for any face motion between frames I and K (920). The face region in frame K is compensated for the computed face motion (930). The compensated region is then subtracted from the corresponding face region in I (940), and the difference value, D(I,K), is used to determine whether K is to be defined as a new member of the Face-Characteristic-View set (950). If the difference between the corresponding faces in the two frames is not large, the next frame K+1 is loaded in the next loop. In case that the difference is in fact larger than a predefined threshold, the face in frame K is input to the set *C*, reference frame I is set to K, and K is set to the next frame K+1 (960). The process, as shown in FIG. 9, is terminated at the end of a face track. At the end of a track, the set *C* contains the most distinctive views of the face.

The set of Face-Characteristic-Views, *C*, is a set of face templates at varying viewpoint angles. Contiguous frames of similar face appearance can be reduced to a single face-frame. Frames that are different enough, in the sense that they can not be reconstructed from existing frames via a similarity

transformation, are included in the set. A future recognition process can therefore be less sensitive to the following (and other) parameters: viewpoint angle of the face; facial expressions, including opening vs closing of the mouth; blinking; and external distracts, such as sunglasses. The Face-Characteristic-View set, *C*, also enables the identification of *dominant features* of the particular face, including (among others): skin-tone color; eye-color; hair shades; any special marks (such as birth marks) that are consistent within the set.

In an anchorperson scene, the set *C* will contain a limited set of views, as there is limited variability in the face and its viewpoint. In a more dynamic scene, the set will contain a large number of entries per face, encompassing the variety of viewpoints of the person in the scene.

The Face-Frontal-View set, *F*, is a set of face templates of the more frontal views of the face. These frames are generally the most-recognizable frames. The selection process is implemented by symmetry-controlled and quality-controlled criteria: In a preferred embodiment the score is computed from correlation values of eyes and mouth candidate regions with at least one eye and mouth template set, respectively. In another preferred embodiment, the quality index depends on a face orientation score computed from a mirrored correlation value of the two eyes. In yet another embodiment, the face centerline is estimated from mouth and nose location, and the face orientation score is computed from the ratio of distances between the left/right eye to the facial centerline. In yet another embodiment, the face quality index includes also a measure of the occlusion of the face in which an approximating ellipse is fitted to the head contour, and the ellipse is tested for intersection with the frame boundaries. In yet another embodiment, the ellipse is tested for intersection with other regions.

The prior art describes a variety of face recognition methods, some based on correlation techniques between input templates, and others utilize distance metrics between feature sets. In order to accommodate the recognition process, a set of Face-Characteristic-Data is extracted. In a preferred embodiment Face-Characteristic-Data include the following:

- Fg = Global information; consists of face templates at selected viewpoints, such as bounding boxes surrounding the face region, as shown in FIG. 10a. Templates of facial components may also be included, in one implementation of eyes, nose and mouth. Templates are known in prior art as one of the means for face recognition.

- Ff = Facial feature geometrical information indicative of the relationships between the facial components. Geometrical information can include interocular distance between the eyes, vertical distance between the eyes and mouth, width of chin, width of mouth and so on; as shown in FIG. 10b. Geometrical features, such as the ones mentioned herein, are known in the prior art as one of the means for face recognition.

- Fu = Unique characteristics, such as eyeglasses, beard, baldness, hair color. Extracting each of these characteristics can be done automatically with algorithms known in the prior art. In one embodiment, following the detection of eyes, a larger window is opened surrounding the eye region, and a search is conducted for the presence of glasses, for example searching for strong edges underneath the eyes, using classification techniques such as neural-networks. A vertical distance can be taken from the eyes to the mouth, and further towards the chin, to check for a beard. In one embodiment the check may be for a color that is different from the natural skin color. In order to extract the characterizing hair color, a vertical projection from the eyes to the forehead can be taken. A rectangular window can be extracted at the suspected forehead and hair line region. A histogram may be used to analyze the color content of the region. If only natural skin color is detected (single mode) there is a strong probability for baldness. In the case of two dominant colors, one is the natural skin color and the second dominant color is the hair color. A probabilistic interpretation can be associated with the results from each of the above-mentioned algorithms. In one embodiment, the associated probabilities can be provided as confidence levels affiliated with each result.

In a preferred embodiment the Face-Characteristic-Data combines the above data, including templates (matrices of intensity values), features vectors containing geometrical information such as characteristic distances or other

information such as coefficients of the eigen-face representation, and a unique characteristic vector, as shown in FIG. 11.

Following the definition of a face segment, audio information in the form of the Audio-Characteristic-Data is incorporated as additional informative characteristic for the segment. One purpose of the present invention is to associate Audio-Characteristic-Data with a face track or part of a face track. Audio-Characteristic -Data may include a set of parameters representing the audio signature of the speaker. In addition, speech-to-text may be incorporated, as known in prior art, to extract recognizable words, and the extracted words can be part of the index.

By combining the results of visual-based face recognition and audio-based speaker identification, the overall recognition accuracy can be improved.

FIG. 12 shows a timeline and video and audio data, which correspond to that timeline. The face/no-face segmentation of the video stream serves as a master temporal segmentation that is applied to the audio stream. The audio segments derived can be used to enhance the recognition capability of any person recognition system built on top of the indexing system constructed according to the present invention.

A further purpose of the present invention is to match audio characteristic data which correspond to two different face tracks in order to confirm the identity of the face tracks. The present invention may utilize prior art methods of audio characterization and speaker segmentation. The latter is required for the case the audio may correspond to at least two speakers.

FIG. 13 shows how to combine the audio track with the face index to create an audio-visual index. The present invention may use prior art methods in speech processing and speaker identification.

It is known in the prior art to model speakers by using Gaussian Mixture Models (GMM) of their acoustic features. A model for a speaker is trained from an utterance of his speech by first computing acoustic features, such as mel-frequency cepstral coefficients, computed every 10ms, and then, considering each feature as a multi-dimensional vector, fitting the parameters of

a mixture of multi-dimensional Gaussian distributions to the said vectors. The set of parameters of the mixture is used to represent the speaker.

Given a speech utterance, it may be scored against a speaker model, as described below, to determine whether the speech belongs to that speaker. First, we derive acoustic features as in the training phase (e.g. computing mel-frequency cepstral coefficients), then, considering each feature as a multi-dimensional vector and viewing the model as a probability distribution on such vectors, we compute the likelihood of these features. In a closed-set setting, a group of speakers is given and a model for each speaker of that group is trained. In this setting, it is a-priori known that the speech utterance belongs to one of the speakers in the group. Thus, computing the likelihood of each speaker model and taking the one with maximum likelihood identifies the correct speaker. In an open-set setting, such a prior knowledge is unavailable. In that case, the speaker verification problem is actually a hypothesis test. The likelihood of the speaker model is compared with the likelihood of so-called cohort model, representing the alternative hypothesis (that the speech utterance belongs to another speaker). If the ratio of the likelihood passes a threshold, the utterance is said to belong to that speaker.

In a preferred embodiment, no prior knowledge of the speakers is assumed. In that embodiment, unsupervised speaker segmentation may be done using an iterative algorithm (1340). Parameters for the speaker GMMs are first initialized using a clustering procedure and than are iteratively improved using the Viterbi algorithm to compute segmentation. Alternatively, one may first detect audio-track changes, or cuts, by considering a sliding window of the audio and testing the hypothesis that this window has no cuts against the hypothesis that there is a cut. Then speaker identification may be performed on each segment (defined as a portion of the audio track between cuts). Audio cut detection may be done by a generalized likelihood ratio test, in which the first hypothesis ("no cuts") is modeled by fitting a best Gaussian distribution to the acoustic feature data derived from the audio window, and the second hypothesis ("there is a cut") is modeled by exhausting over cut points, fitting a best Gaussian to each part of the window (left or right to the cut point), and

taking the cut point for which the likelihood is maximal. The likelihood of the first hypothesis is compared to the likelihood of the second hypothesis, and a threshold is used in the decision process (taking into account the larger number of parameters in the second hypothesis).

FIG. 13 illustrates a preferred embodiment, wherein the segmentation of the audio track is aided by visual cues from the video face indexing process (1310). In particular, the audio is partitioned with respect to the video content. For example when an entire shot includes a single face track, the initial hypothesis can be a single speaker. Once verified, the audio characteristic data (such as the GMMs parameters) are associated with that face (1350). In another example, when an entire shot includes only two faces, the initial hypothesis can be two speakers. Once verified, the audio characteristic data of a speech segment are associated with the face of highest mouth activity as computed by the visual mouth activity detector (1320). Tracking mouth activity is known in prior art. Such technology may entail a camera capturing the mouth image followed by thresholding the image into two levels, i.e., black and white. Binary mouth images are analyzed to derive the mouth open area, the perimeter, the height, and the width. These parameters are then used for recognition. Alternatively, the binary mouth images themselves may be used as the visual feature, with clustering schemes used to classify between these binary images. Derivative information of the mouth geometrical quantities, as well as optical flow input have been suggested as well. Prior art also entails the detection of the lip boundaries (upper and lower lips), the modeling of the lip boundaries via splines and the tracking of the spline movements across frames. Motion analysis of the lip boundaries (spline models) can facilitate an overall analysis of visual mouth activity, such as the segmentation into talking vs. non-talking segments. Further analysis utilizing learning technologies such as neural-networks, and additional techniques such as a technique called "Manifold Learning", can facilitate learning complex lip configurations. These techniques are used in the prior art to augment speech recognition performance. In this invention they are utilized for the detection of motion and activity towards the affiliation of an audio signature to a visual face, as well as the affiliation of

extracted words from the audio to the visual face.

Once the audio track is segmented (1330,1340) and associated with a video track (1350), audio characteristic data from the speech segment can augment the face index. In a preferred embodiment as described in Figure 14, a speech to text engine is employed (1410) to generate a text transcription of the speech segments in the video program. Such engines are commercially available [e.g. from Entropic Ltd. from Cambridge, England; or ViaVoice from IBM]. Such an engine may operate on a computer file representation of the audio track or in real time using the audio signal input into a computer with audio digital card (such as Sound Blaster by Creative Labs). A full transcription of the audio may be extracted by speech recognition technology, as described above. In addition, closed-caption decoding may be utilized, when available. In one embodiment, full transcription can be incorporated. In yet another embodiment, a subset of predefined keywords may be extracted. The transcription of the audio is next aligned with the video face index. The alignment is achieved by means of the related time codes (1420). In one embodiment said alignment may include the identification of each word or utterance start and end points, hereby termed the word time segment. A word is associated with a face track for which there is an overlap between the word time segment and the face track time segment. In addition, as described above for the case of attaching audio signature data for speaker identification and person recognition, the transcription may be attached as text to the face track that is determined visually to be the most matching, what we may term as speech to speaker matching. For example when an entire shot includes a single face track, the initial hypothesis can be a single speaker. The transcription is attached in full to the matching face. In another example, when an entire shot includes two faces, the audio transcription of a speech segment is associated with the face of highest mouth activity as computed by the visual mouth activity detector (1320).

In one embodiment of the invention, there is a need to identify face segments or part of face segments, as talking segments or non-talking segments, in which the associated face is talking or not-talking, respectively.

This information is valuable information to be included in the face index store, enabling future search for a person in a talking state. Moreover, this information may be critical for the matching process described above between speech and speaker. In a preferred embodiment, in order to partition the face segments into talking and non-talking segments, mouth motion is extracted in a face track, tracking the mouth in a frame-by-frame basis throughout the track. The extracted motion characteristics are compared against talking lips movement characteristics and a characterization is made into one of two states: talking vs. non-talking.

According to the present invention the extracted information from a face track may be incorporated into the face index, and links may be provided between similar face tracks to merge the information. The linking and information-merging stage, as part of the face indexing process, is depicted in Fig. 15. If the face index store is empty (1510), the current face track initializes the index, providing its first entry. Otherwise, distances are calculated between the new face track characteristics and the characteristics of each face entry in the index (1520). In a preferred embodiment, distances can be computed between several fields of the Face-Characteristic-Data. Template data (Fg in FIG 11), can be compared with correlation-based techniques. Feature vector data (Ff in FIG 11), that represent geometrical characteristics or eigen-basis characteristics, can be compared utilizing distance metrics as known in the art, such as the Euclidean metric or the Mahalanobis metric. Similarly with the Unique characteristic data (Fu in FIG 11). Overall distance measure is calculated as a function of the individual distance components, in one embodiment being the weighted sum of the distances. Distance measures are ranked in increasing order (1530). The smallest distance is compared to a *Similarity threshold* parameter (1540) to categorize the entry as a new face to be entered to the index, or as an already existing face, in which case the information is merged to an existing entry in the index.

The description has so far focused on the generation of the face index (FIG. 3), involving the processing of the inputted video and the grouping of the information into tracks and the tracks into appropriate index entries. As part of

the information merging process, an additional grouping step can be considered to better facilitate browsing and searching and augment recognition performance. In one embodiment, once the index contains multiple track entries per index entry, and at predefined time intervals, the information associated with the index entry from all corresponding tracks can be merged, to produce representative set of characteristic data. In a preferred embodiment, such a set involves the best quality frontal face view (ranking frontal face views from all tracks); a second set includes the conjunction of all the unique characteristic data.

In a preferred embodiment, the editor can annotate the face index. Annotation may be implemented in several modes of operation. In one, a description or label may be attached manually to an entry in the index, following which the annotation becomes automatically linked to all frames in all tracks that relate to the index entry. In a second mode of operation, following the annotation process, each new frame that is added incrementally to a respective location in the index store, automatically receives all annotations available for the respective index. Annotation descriptions or labels may be inputted manually by an operator (270). Alternatively, any characteristic data in the data store (e.g. the occurrence of glasses, beard) may be automatically converted, by means of a stored dictionary, to a text description or label and added on the annotation fields. In this fashion, attributes, features and names, in a textual form, are slowly incorporated into the index, automatically, enabling text search and increasing search speeds. In yet an additional mode of operation, audio characteristic data, such as keywords detected within a segment, or the entire spoken text, can be converted via speech-to-text, as known in the art, to incorporate annotation of audio information as additional information in the index.

The annotation process as described above serves to augment the generated index as taught by the present invention. The present invention further teaches how to generate video annotation from the index annotation, as depicted in 280 of FIG 2. In one embodiment each video frame may have associated with it a text string that contains the annotations as present in the

related index entry. A video annotation attached to each video frame is exemplified in FIG 16. The index annotation field of FIG. 5 (right-most field) is now linked to the related video frames, as shown in FIG. 16. A single index entry is associated with a video track, and a video track may be associated in turn with a large number of video frames. Each frame in the corresponding sequence is now (automatically) annotated. Note that the annotation may include a label (such as the person name), any manually inputted description (such as the affiliation of the person), any description of visual features (e.g., 'beard', 'glasses'), and other, as taught in this invention. In another embodiment, a listing may be generated and displayed, as shown in Table 1, summarizing all annotated information as related to frame segments.

Start Frame	End Frame	Annotation
Fs	Fe	"Clinton"

Table 1

The face index store is used for browsing and searching applications. A user interested in browsing can select to browse at varying resolutions. In one browsing mode, the display includes the best quality- most frontal face template (single image of face on screen). In a second browsing mode, the user may select to see the set of best characteristic views (e.g. one row of face images of distinct views). In a third browsing mode, user may select to see all views in all tracks available in the store for the person of interest. Several screens of frames will be displayed. Frames may be ordered based on quality metrics, or they may

be ordered via a time line per the video source. The user may interactively select the mode of interest and shift amongst them.

The present invention teaches a method of searching in video (or in a large collection of images) for particular members of an object class, by generating the above-described index and searching the index. FIG. 17 presents a flowchart of the searching methodology. The input search query may address any of the variety of fields present in the index, and any combinations therefrom. Corresponding fields in the index will be searched (1710). The user can search for example by the name of the person. The label or description inputted by the user is compared to all labeled data; if a name in fact exists in the index, all related face tracks would be provided as search results. The user can also search by providing an image of a face and searching for similar looking faces; in this case, template data will be used, such as the best quality face template, or the best set of characteristic views per person entry. The user can also search by attribute data; in this scenario, feature vectors are compared. The user can query using multiple attributes, such as the face label as well as a set of keywords; in this case, the search would be conducted simultaneously on multiple parameters, such as the face template and audio characteristics.

Matching index entries are extracted and can be optionally ranked and sorted based on the match quality (1720). A match score may be computed by correlating the input template to the best quality template per index store entry; or the input template can be compared with the best set of characteristic view templates; or the input characteristic data can be compared with stored characteristics per index store entry, etc. In any comparison mode, the final match score can be chosen as the best match score, or as a weighted combination of the scores.

The matched index entries are automatically associated with the corresponding sets of video frames (1730).

The search results may be displayed in a variety of forms. In one embodiment a listing of the index entries may be outputted. In another embodiment, the video frames are displayed as icons. Output frames may be

ordered based on match quality; alternatively they may be displayed on a time-line basis per related video source (e.g. per tape or movie source).

Utilizing the face index store allows for much greater efficiency in browsing and searching of the video data, as well as increasing recognition accuracy. Automatic organization of the data into representative sets, merged across multiple tracks from different video sources, allows for efficient browsing, at varying resolutions, as described above. In the searching mode, rather than searching on a frame-by-frame basis for the person of interest, a comparison is made on the indexed material in the face index store.

The foregoing description has been concerned primarily with the indexing of facial content in video sequence. However, the methods taught by the invention can be readily applied to include indexing other classes of objects for browsing and searching purposes.